REMARKS

The application is believed to be in condition for allowance because the claims are novel and non-obvious over the cited art. The following paragraphs provide the justification for these beliefs. In view of the following reasoning for allowance, the applicants hereby respectfully request further examination and reconsideration of the subject application.

Claim Objections

The applicants acknowledge the Examiner's claim renumbering and have rewritten the claims to reflect the correct dependency. It is believe that these amendments have overcome any objections to the claims.

The 35 USC 103 Rejection of Claims 1-2, 9-10, 12, 24 and 18-31.

Claims 1-2, 9-10, 12, 24 and 18-31 were rejected under 35 USC 103(a) as being unpatentable over Maali et al, U.S. Patent No. 6,567,775, herein after referred to as Maali, in view of Stork, U.S. Patent No. 5,586,215 (herein after Stork). The Examiner stated that Maali teaches the applicants' claimed invention, but does not teach using audio and video signals to train a time delay neural network to determine when a person is speaking. However, the Examiner further contended that Stork teaches this feature, rendering the applicants' claimed invention obvious. The applicants respectfully traverse this contention of obviousness.

In order to deem the applicants' claimed invention unpatentable under 35 USC 103, a prima facie showing of obviousness must be made. To make a prima facie showing of obviousness, all of the claimed elements of an applicant's invention must be considered, especially when they are missing from the prior art. If a claimed element is not taught in the prior art and has advantages not appreciated by the prior art, then no prima facie case of obviousness exists. The Federal Circuit court has stated that it was error not to distinguish claims over a combination of prior art references where a

material limitation in the claimed system and its purpose was not taught therein (*In Re Fine*, 837 F.2d 107, 5 USPQ2d 1596 (Fed. Cir. 1988)).

The applicants claim a system and method for utilizing the correlation between video and audio input from a single microphone to detect speakers. A time-delayed neural network (TDNN) is trained to learn the audio-visual correlation in speaking. This trained TDNN is then used to search one or more audio-video inputs to detect when a person in the audio-video input is speaking. The audio-video speaker detection system of the applicant's claimed invention overcomes limitations in previous speaker detection system in that it has an improved TDNN structure, better video features, and simpler audio features. The speaker detection system and method of the invention employs a face detector to accurately locate the faces of speakers, and compensates for head motion by stabilizing mouth location.

In one aspect, the applicant's claimed invention collects audio data generated from a microphone device. In another aspect, the applicant's claimed invention collects video data and processes the data to determine a mouth location for a given speaker. The audio and video are input into a TDNN that processes the data to determine when a given person is speaking. The neural network processing is based upon a correlation to detected mouth movement from the video data and audio sounds detected by the microphone. (Summary)

The audio-video speaker detection system and method according to the applicant's claimed invention computes audio and video features. In one embodiment of the invention, the audio feature is the energy over an audio frame (an audio is typically 33 ms for 30 FPS video) that is useful as it is the simplest possible, and is only one-dimensional. (Summary)

The video features are determined by first using a face detector to locate a face in an audio video clip. Using the geometry of a typical face, the mouth location is then estimated. In one embodiment of the applicant's claimed invention, the mouth image sequence is then stabilized using normalized correlation to remove any translational motion of the mouth caused by head movement. However, other

stabilization techniques could be used to stabilize the mouth region. Finally, a Linear Discriminant Analysis (LDA) projection is used to determine if the mouth in the mouth image is open or closed. Based on the LDA projections, the values of the mouth openness for each segment of audio video data is designated in values ranging from -1 for the mouth being closed, to +1 for the mouth being open. (Summary)

These audio and video features are then used to train the TDNN to recognize when a person in an input audio video clip is speaking. Input **audio and video signals are used to train a time delay neural network to determine when a person is speaking by 1) computing audio features from said audio training data wherein the audio feature is the energy over an audio frame; 2) computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed; and 3) correlating said audio features and video features to determine when a person is speaking.**

Once the TDNN is trained, the trained TDNN is used to determine if a detected speaker in an audio-video sequence is speaking.

In contrast, Maali discloses a method and apparatus for identifying a speaker. An audio-based speaker identification system identifies one or more potential speakers for a given segment using an enrolled speaker database. A video-based speaker identification system identifies one or more potential speakers for a given segment using a face detector/recognizer and an enrolled face database. An audio-video decision fusion process evaluates the individuals identified by the audio-based and video-based speaker identification systems and determines the speaker of an utterance. **A linear variation is imposed on ranked-lists produced using the audio and video information. <u>The decision fusion scheme of the Maali invention is based on a linear combination of the audio and the video ranked-lists.</u> The line with the higher slope is assumed to convey more discriminative information. The normalized slopes of the two lines are used as the weight of**

the respective results when combining the scores from the audio-based and video-based speaker analysis. In this manner, the weights are derived from the data itself. (Abstract)

Maali does not, however, teach the applicant's claimed using audio and video signals to train a time delay neural network to determine when a person is speaking, wherein the training comprises...1) computing audio features from said audio training data wherein the audio feature is the energy over an audio frame; 2) computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed; and 3) correlating said audio features and video features to determine when a person is speaking.

Granted, the Examiner states that Maali teaches computing audio features wherein the audio feature is the energy over an audio frame at column 3 lines 51-60 with column 6, lines 5-24), but the first passage merely teaches inputting video and audio data with multiple speakers. Column 3, lines 51-60 teaches using cepstral features which are computed using the signal energy in various frequency bands, not the entire audio frame. Additionally, both the audio and video features used in Maali are not used for training any type of Neural Network. In fact, Maali does not even employ a Neural Network, much less train one. **Additionally, Maali does not correlate the audio and video features but employs a decision fusion scheme based on a linear combination of the audio and the video ranked-lists.**

Stork teaches a neural network acoustic and visual speech recognition system for the recognition of speech comprises an acoustic preprocessor, a visual preprocessor, and a speech classifier that operates on the acoustic and visual preprocessed data. The acoustic preprocessor comprises a log mel spectrum analyzer that produces an equal mel bandwidth log power spectrum. The visual processor detects the motion of a set of fiducial markers on the speaker's face and extracts a set of normalized distance vectors describing lip and mouth movement.

The speech classifier uses a multilevel time-delay neural network operating on the preprocessed acoustic and visual data to form an output probability distribution that indicates the probability of each candidate utterance having been spoken, based on the acoustic and visual data. Stork does not determine when a person is speaking, but determines which of a given set of utterances a person is speaking.

Additionally, Stork does not teach the applicant's claimed using audio and video signals to train a time delay neural network to determine when a person is speaking, wherein the training comprises...1) computing audio features from said audio training data wherein the audio feature is the energy over an audio frame; 2) computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed; and 3) correlating said audio features and video features to determine when a person is speaking.

Since neither Maali nor Stork teaches the applicant's claimed **using audio and video signals to train a time delay neural network to determine when a person is speaking, wherein the training comprises...1) computing audio features from said audio training data wherein the audio feature is the energy over an audio frame; 2) computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed; and 3) correlating said audio features and video features to determine when a person is speaking,** the combination does not teach it. Thus, the applicants have claimed elements not taught in the cited art and which have advantages not recognized therein. Namely since the applicants' claimed invention uses low-level correlation of audio/video to detect speakers, the accuracy of speaker detection is better than using audio alone (e.g., with a microphone array) or even high-level audio/video fusion. (Summary) Accordingly, no prima facie case of obviousness has been established in accordance with the holding of *In Re Fine*. This lack of prima facie showing of obviousness means that the rejected claims are

patentable under 35 USC 103 over Maali in view of Stork. It is, therefore, respectfully requested that the rejection of Claims 1-2, 9-10, 12, 24 and 18-31 be reconsidered based on the novel and non-obvious claim language:

> "using said audio and video signals to train a time delay neural network to determine when a person is speaking, wherein said training comprises the following process actions....**computing audio features from said audio training data wherein said audio feature is the energy over an audio frame**; computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed; **and correlating said audio features and video features to determine when a person is speaking."**

The 35 USC 103 Rejection of Claims 3-5 and 11.

Claims 3-5 and 11 were rejected under 35 USC 103(a) as being unpatentable over Maali in view of Stork, and in further view of Nefian et al., U.S. Patent No. PGPUB2004/0122675 (herein after Nefian). The Examiner stated that Maali and Stork teach the applicants' claimed invention, but do not teach reducing the noise of the audio signals during preprocessing. However, the Examiner further contended that Nefian teaches this feature, rendering the applicants' claimed invention obvious. The applicants respectfully disagree with this contention of obviousness.

As discussed above, the applicants claim a technique for detecting speech. Associated audio and video training data containing a person's face that is periodically speaking are used to train a time delay neural network to determine when a person is speaking. The training comprises: 1) computing audio features from the audio training data wherein said audio feature is the energy over an audio frame; 2) computing video features from the video training signals wherein the video feature is the degree to which said person's mouth is open or closed; and 3) correlating the audio features and video features to determine when a person is speaking.

As discussed above, neither Maali nor Stork teaches the applicant's claimed **using audio and video signals to train a time delay neural network to determine when a person is speaking, wherein the training comprises...1)**

computing audio features from said audio training data wherein the audio feature is the energy over an audio frame; 2) computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed; and 3) correlating said audio features and video features to determine when a person is speaking.

Nefian et al teaches a speech recognition method that includes several embodiments of applying support vector machine analysis to a mouth region. Lip position can be accurately determined and used in conjunction with synchronous or asynchronous audio data to enhance speech recognition probabilities. (Audio) But Nefian does not teach the applicant's claimed **using audio and video signals to train a time delay neural network to determine when a person is speaking, wherein the training comprises...1) computing audio features from said audio training data wherein the audio feature is the energy over an audio frame; 2) computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed; and 3) correlating said audio features and video features to determine when a person is speaking.**

Additionally, the Examiner states that Nefian teaches reducing the noise of the audio signals at paragraph 26. This passage reads,

> "The audio and video observation likelihoods are computed independently, significantly reducing the parameter space and overall complexity of the Markov model. Viseme/phoneme pairs are modeled, and in certain embodiments, **audio and visual probabilities** can be modified to adjust for differing associated noise levels."

This passage does not state that the noise of the audio signals is reduced during preprocessing. It merely states the audio and visual probabilities can be modified to adjust for differing associated noise levels.

Since neither Maali nor Stork nor Nefian teaches the applicant's claimed **using audio and video signals to train a time delay neural network to determine when a person is speaking, wherein the training comprises...1) computing audio features from said audio training data wherein the audio feature is the energy over an audio frame; 2) computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed; and 3) correlating said audio features and video features to determine when a person is speaking,** the combination does not teach it. Additionally, Nefian does not teach reducing noise of the audio signals during preprocessing. Thus, the applicants have claimed elements not taught in the cited art and which have advantages not recognized therein. Accordingly, no prima facie case of obviousness has been established in accordance with the holding of *In Re Fine*. This lack of prima facie showing of obviousness means that the rejected claims are patentable under 35 USC 103 over Maali in view of Stork and Nefian. It is, therefore, respectfully requested that the rejection of Claims 3-5 and 11 be reconsidered based on above-quoted novel and non-obvious claim language.

The 35 USC 103 Rejection of Claims 13-15 and 20-22.

Claims 13-15 and 20-22 were rejected under 35 USC 103(a) as being unpatentable over Bakis et al , U.S. Patent No. 6,219,639 (hereinafter Bakis) in view of Stork. The Examiner stated that Bakis teaches the applicants' claimed invention, but does not teach training a Time Delay Neural Network to determine when a person is peaking using extracted audio and video features. However, the Examiner further contended that Stork teaches this feature, rendering the applicants' claimed invention obvious. The applicants respectfully disagree with this contention of obviousness.

Bakis teaches a method for recognizing an individual based on attributes associated with the individual. The method comprises the steps of: pre-storing at least two distinctive attributes of the individual during at least one enrollment session; contemporaneously extracting the at least two distinctive attributes from the

individual during a common recognition session; segmenting the pre-stored attributes and the extracted attributes according to a sequence of segmentation units; indexing the segmented pre-stored and extracted attributes so that the segmented pre-stored and extracted attributes corresponding to an identical segmentation unit in the sequence of segmentation units are associated to an identical index; and respectively comparing the segmented pre-stored and extracted attributes associated to the identical index to each other to recognize the individual. (Abstract).

But Bakis does not teach the applicant's claimed **using audio and video signals to train a time delay neural network to determine when a person is speaking, wherein the training comprises...1) computing audio features from said audio training data wherein <u>the audio feature is the energy over an audio frame</u>; 2) computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed; and 3) <u>correlating said audio features and video features to determine when a person is speaking.</u>**

The Examiner states that Bakis teaches extracting audio features wherein an audio features is the energy over an audio frame and wherein said video feature is the openness of a person's mouth at column 10, lines 5-35. However, this passage says nothing about the audio feature being the energy over an audio frame. In no way does Bakis teach this feature of the applicant's claimed invention.

Stork also does not teach the applicant's claimed **using audio and video signals to train a time delay neural network to determine when a person is speaking, wherein the training comprises...1) computing audio features from said audio training data wherein <u>the audio feature is the energy over an audio frame</u>; 2) computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed; and 3) <u>correlating said audio features and video features to determine when a person is speaking.</u>** Additionally, Stork does not teach training a Time Delay Neural Network to determine when a person is speaking using the extracted

audio and video features. Stork teaches training a Neural Network determine the **probability of an utterance being one of the utterance that it has been trained to recognize. Therefore, Stork teaches training of a neural network to recognize <u>what</u> is said, not to recognize <u>when</u> someone is talking.**

Since neither Bakis nor Stork teaches the applicant's claimed **using audio and video signals to train a time delay neural network to determine when a person is speaking, wherein the training comprises...1) computing audio features from said audio training data wherein <u>the audio feature is the energy over an audio frame</u>; 2) computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed; and 3) <u>correlating said audio features and video features to determine when a person is speaking,</u>** the combination does not teach it. Thus, the applicants have claimed elements not taught in the cited art and which have advantages not recognized therein. Accordingly, no prima facie case of obviousness has been established in accordance with the holding of *In Re Fine*. This lack of prima facie showing of obviousness means that the rejected claims are patentable under 35 USC 103 over Bakis in view of Stork. It is, therefore, respectfully requested that the rejection of Claims 13-15 and 20-22 be reconsidered based on above quoted novel and non-obvious claim language.

<u>The 35 USC 103 Rejection of Claim 16.</u>

Claim 16 was rejected under 35 USC 103(a) as being unpatentable over Bakis in view of Stork and in further view of Nefian. The Examiner stated that Bakis and Stork teach the applicants' claimed invention, but do not teach an instruction for reducing noise in the audio video clips prior to segmenting them. However, the Examiner further contended that Nefian teaches this feature, rendering the applicants' claimed invention obvious. The applicants respectfully disagree with this contention of obviousness.

As discussed above, Bakis and Stork and Nefian do not teach the applicant's claimed using audio and video signals to train a time delay neural network to determine when a person is speaking, wherein the training comprises...1) computing audio features from said audio training data wherein the audio feature is the energy over an audio frame; 2) computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed; and 3) correlating said audio features and video features to determine when a person is speaking.

Additionally, as discussed previously, Nefian does not teach that the noise of the audio signals is reduced during preprocessing. It merely states the audio and visual probabilities that it employs can be modified to adjust for differing associated noise levels.

Since neither Bakis nor Stork nor Nefian teach the applicant's claimed **using audio and video signals to train a time delay neural network to determine when a person is speaking, wherein the training comprises...1) computing audio features from said audio training data wherein the audio feature is the energy over an audio frame; 2) computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed; and 3) correlating said audio features and video features to determine when a person is speaking,** the combination does not teach it. Additionally, Nefian does not teach reducing the noise of the audio signals during preprocessing. Thus, the applicants have claimed elements not taught in the cited art and which have advantages not recognized therein. Accordingly, no prima facie case of obviousness has been established in accordance with the holding of *In Re Fine*. This lack of prima facie showing of obviousness means that the rejected claims are patentable under 35 USC 103 over Bakis in view of Stork and in further view of Nefian. It is, therefore, respectfully requested that the rejection of Claim 16 reconsidered based on above quoted novel and non-obvious claim language.

The 35 USC 103 Rejection of Claims 17-18 and 23.

Claims 17-18 and 23 were rejected under 35 USC 103(a) as being unpatentable over Bakis in view of Stork and in further view of Liang, US Publication Number 2003/0212552 (herein after Liang). The Examiner stated that Bakis and Stork teach the applicants' claimed invention, but do not teach stabilizing the mouth using Linear Discriminant Analysis and designating values for the mouth. However, the Examiner further contended that Liang teaches this feature, rendering the applicants' claimed invention obvious. The applicants respectfully disagree with this contention of obviousness.

As discussed above, Bakis and Stork do not teach the applicant's claimed using audio and video signals to train a time delay neural network to determine when a person is speaking, wherein the training comprises...1) computing audio features from said audio training data wherein the audio feature is the energy over an audio frame; 2) computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed; and 3) correlating said audio features and video features to determine when a person is speaking.

Liang teaches a face recognition procedure useful for audiovisual speech recognition. A visual feature extraction method includes application of multiclass linear discriminant analysis to the mouth region. However, Liang does not teach the applicant's claimed using audio and video signals to train a time delay neural network to determine when a person is speaking, wherein the training comprises...1) computing audio features from said audio training data wherein the audio feature is the energy over an audio frame; 2) computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed; and 3) correlating said audio features and video features to determine when a person is speaking.

Since neither Bakis nor Stork nor Liang teach the applicant's claimed **using audio and video signals to train a time delay neural network to determine when**

a person is speaking, wherein the training comprises...1) computing audio features from said audio training data wherein <u>the audio feature is the energy over an audio frame</u>; 2) computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed; and 3) <u>correlating said audio features and video features to determine when a person is speaking,</u> the combination does not teach it. Thus, the applicants have claimed elements not taught in the cited art and which have advantages not recognized therein. Accordingly, no prima facie case of obviousness has been established in accordance with the holding of *In Re Fine.* This lack of prima facie showing of obviousness means that the rejected claims are patentable under 35 USC 103 over Bakis in view of Stork and in further view of Liang. It is, therefore, respectfully requested that the rejection of Claims 17-18 and 23 reconsidered based on above quoted novel and non-obvious claim language.

<u>The 35 USC 103 Rejection of Claims 25-27.</u>

Claims 17-18 and 23 were rejected under 35 USC 103(a) as being unpatentable over Maali in view of Stork and in further view of Liang, and in further view of PGPUB 2004/0267521. This publication is the publication of the current patent application, so the applicant does not understand how it can be cited as prior art. Regardless, the Examiner stated that Maali and Stork and Liang teach the applicants' claimed invention, but do not teach using a microphone array beam form on each face. However, the Examiner further contended that beamforming is well known to improve the sound quality of a speaker. The applicants respectfully disagree with this contention of obviousness.

As discussed above, Bakis and Stork and Liang do not teach the applicant's claimed using audio and video signals to train a time delay neural network to determine when a person is speaking, wherein the training comprises...1) computing audio features from said audio training data wherein <u>the audio feature is the energy over an audio frame</u>; 2) computing video features from said video training signals wherein said video feature is the degree to which said person's

mouth is open or closed; and 3) correlating said audio features and video features to determine when a person is speaking.

Furthermore, while it may be well known to use beamforming to improve the audio quality of a speaker, this does not mean that it is well known to compute video features from said video training signals by using a face detector to locate each face in said video training signals and using a microphone array to beam form on each face detected thereby filtering out sound not coming from the direction of the speaker to create beam formed audio training data. The applicant certainly did not admit that this was well known, and none of the cited references teach this feature.

Furthermore, since neither Bakis nor Stork nor Liang teach the applicant's claimed **using audio and video signals to train a time delay neural network to determine when a person is speaking, wherein the training comprises...1) computing audio features from said audio training data wherein the audio feature is the energy over an audio frame; 2) computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed; and 3) correlating said audio features and video features to determine when a person is speaking,** the combination does not teach it. Thus, the applicants have claimed elements not taught in the cited art and which have advantages not recognized therein. Accordingly, no prima facie case of obviousness has been established in accordance with the holding of *In Re Fine*. This lack of prima facie showing of obviousness means that the rejected claims are patentable under 35 USC 103 over Bakis in view of Stork and in further view of Liang. It is, therefore, respectfully requested that the rejection of Claims 25-27 reconsidered based on above quoted novel and non-obvious claim language.

The 35 USC 103 Rejection of Claim 32.

Claim 32 was rejected under 35 USC 103(a) as being unpatentable over Maali in view of Stork and in further view of Nefian. The Examiner stated that Maali and Stork teach the applicants' claimed invention, but do not teach a statistical learning engine

that is a Support Vector Machine. However, the Examiner further contended that Nefian teaches this feature, rendering the applicants' claimed invention obvious. The applicants respectfully disagree with this contention of obviousness.

As discussed above, Maali and Stork and Nefian do not teach the applicant's claimed using audio and video signals to train a time delay neural network to determine when a person is speaking, wherein the training comprises...1) computing audio features from said audio training data wherein <u>the audio feature is the energy over an audio frame</u>; 2) computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed; and 3) <u>correlating said audio features and video features to determine when a person is speaking.</u>

Since neither Maali nor Stork nor Nefian teach the applicant's claimed **using audio and video signals to train a time delay neural network to determine when a person is speaking, wherein the training comprises...1) computing audio features from said audio training data wherein <u>the audio feature is the energy over an audio frame</u>; 2) computing video features from said video training signals wherein said video feature is the degree to which said person's mouth is open or closed; and 3) <u>correlating said audio features and video features to determine when a person is speaking,</u>** the combination does not teach it. Thus, the applicants have claimed elements not taught in the cited art and which have advantages not recognized therein. Accordingly, no prima facie case of obviousness has been established in accordance with the holding of *In Re Fine*. This lack of prima facie showing of obviousness means that the rejected claims are patentable under 35 USC 103 over Maali in view of Stork and in further view of Nefian. It is, therefore, respectfully requested that the rejection of Claim 32 reconsidered based on above quoted novel and non-obvious claim language.

Allowable Subject Matter

The applicants gratefully acknowledge the allowability of Claims 8 and 19 if rewritten in independent form including all of the limitation of the base claim and any intervening claims. The applicants, at this time, decline to rewrite these claims in independent form at it is believed that the claims as written are patentable over the cited art.

In summary, it is believed that claims 1-28 and 31-32 are in condition for allowance. Allowance of these claims at an early date is courteously solicited.

LYON & HARR, LLP
300 Esplanade Drive
Suite 800
Oxnard, CA 93036
(805) 278-8855

Respectfully submitted,

Katrina A. Lyon
Reg. No. 42,821
Attorney for Applicant(s)